Final Report of PLC Projects

November 23

2009

The projects done in Phase II are: Khmer Sort Message Service, Khmer Typing Game, Khmer Optical Character Recognition (Khmer OCR) able to recognize two Fonts, Improvement on the projects in Phase I such as: Khmer Conversion, Khmer Line Breaking, Khmer Collation, Khmer Spell Checker, Font Creation

Phase II

Contents

I. INTRODUCTION

Cambodia is a developing country where Information and Communication Technologies are still far behind those of their neighboring countries. Information Communication Technology factor is crucial for both enhancing the human capacity and the economics in Cambodia.

Education factor after high school level in Cambodia is still based on the second-language like French or English, especially new-technology-related fields. Moreover, very small number of female students can attend the university study life. These make latency in human resource capacity strengthening, slowing down the economic growth.

PAN Localization Cambodia is a local project funded by International Research Centre of Canada since 2005. The main goal of the project is to take part in the development in Cambodia including the following goals:

- To build sustainable human resource capacity in Cambodia for R&D in Local Language Technology.
- To raise current levels of information technology support for Khmer language in ICT use in Cambodia.
- To promote policy in Cambodia for local language content creation and access, for development;
- To investigate effective, cohesive mechanism to develop and diffuse local language computing capacity in Cambodia.

To reach above goals, PLC team has developed some useful Khmer driven applications such as Khmer Short Message Application, Khmer Typing Application, Khmer Optical Character Recognition, Khmer Keyboard Layout, Khmer Conversion, Khmer Line Breaking, Khmer Sorting, Khmer Spell Checker, Khmer Fonts, and Khmer Text To Speech.

Khmer Unicode has been accepted internationally in ICT field, marking the starting point for Khmer Language to step into Information Technology sector. However, Khmer Unicode alone cannot enrich Khmer language in Computer. The development of Khmer driven applications helps a lot in this case, for example, Khmer Keyboard Layout help enable end-users to input Khmer characters in form of Unicode. Khmer Unicode fonts have been created, allowing end-users to be able to choose their styles of writing. Seeing its advantages both female and male end-users, after a short complaining on changing from their old habit to adapt with the world of Unicode, accept and are proud to type their document based on Khmer Unicode.

Moreover, Khmer Typing Application has been created to answer the difficulties adapting with Khmer Unicode. With the application, both female and male users can learn to type Khmer Unicode based on Keyboard Layout created with some lessons ported with the application. Khmer Keyboard Layout, Khmer Conversion, Khmer Line Breaking, Khmer Sorting, Khmer Spell Checker, Khmer Fonts have been packed into one package for end-user to install in their application and practice in their daily work.

Khmer Short Message Service has been created, but not yet released because we have to remove existing bugs. However, we expect a remarkable behavior change among both female and male end-users on ICT access. Previously, SMS in English, French, or other non-native language have been very popular among Cambodian people, incredibly teenagers and those who get educated higher than secondary school. After the launch of KSMS, we strongly

expect that the end-users will be gradually happy to use the applications as their medium of communication, especially women and teenagers who do not know other foreign languages.

On the other hand, Khmer Optical Character Recognition plays an important role on the young Cambodian fellows Research and Development capacity building. With Khmer OCR, moving the Khmer contents in the website or other application in Computer will be no longer a time consuming task, searching the contents in Khmer will response in an acceptable number of information. Unfortunately, Khmer OCR is able to recognize initially two fonts, so converting document into an editable format is still limited. We are looking ahead to improve and broaden the scope of the application.

Khmer Text To Speech, converting Khmer text into voice, also plays a crucial role in R&D field. The presentation on Khmer TTS attracts high attention among the students to get closer in the Khmer Unicode in ICT field.

In short, the applications developed by PLC involved in development of the country. The works were done by young Cambodian fellows, improving their skills in ICT field. With this fruitful result, Khmer ICT is stepping a head toward the new and up to date technology. We are willing and trying our best to keep on the work for the sake of the development of the country.

II. KHMER LANGUAGE OVERVIEW

Khmer is one of the ancient languages in Asia; its script was inherited from the Indian script. It is influenced by Pali and Sanscrit which are considerably applied mainly among royal and religious by Buddhism and Hinduism. Khmer is a non-tonal language; tonation in words does not change their meaning.

Khmer as the official language of the Kingdom of Cambodia is used in both speaking and writing system. Its writing system differs from those of other languages such as French, English, or Japanese, etc. A word can be formed by a consonant, consonants and vowels, or only an independent vowel. There is neither conjugation of the main verbs nor modification of the main verbs in its grammar rules. Khmer is written continuously without space to delimit the word in the sentence.

Localizing language in the computer is based on the rules and definition of the words in Khmer CHOUN NATH Dictionary which is recognized as the only official dictionary in Cambodia [4]. Khmer has many ambiguities of word meanings and combinations according to their context of the sentence: the variation of their Part of Speech in different contexts. Besides, some words can be joined together and have another meaning and Part Of speech (PoS).

For example, in Table 1, a Khmer sentence ឪពុកខ្ញុំទៅធ្វើការ, which means correctly in English “My father goes to work,” can have a possible word segmentation and Part Of Speech (PoS). The word ពុក can have two meanings “Father,” or “Decay” and PoSes “Noun,” or “Adjective,” respectively. However, the combination of ឪ and ពុក forms ឪពុក which means “Father” with “Noun” as its PoS here is the correct meaning in the context.

Example of a Khmer sentence	
Khmer	ឪពុកខ្ញុំទៅធ្វើការ

English	My father goes to work					
Segmentation	ឪ	ពុក	ខ្ញុំ	ទៅ	ធ្វើ	ការ
Meaning	Father	Father|Decay	Me|I|My	Go	Do	Job/Work
PoS	N	N|Adj	PRP|PRP|RPN	V	V	N
Correct Segmentation	ឪពុក		ខ្ញុំ	ទៅ	ធ្វើការ	
Correct Meaning	Father		My	Go	Work	
Correct PoS	N		RPN	V	V	

Table 1: Khmer segmentation to form a correct meaning

Khmer Script Shape Deformation and Position Changing occur when there is a combination of the consonant and vowel. Table 2 demonstrates how Khmer scripts change their shape in word formation.

Shape Deformation and Position changing				
Input types	Input Shapes			Output
Consonant + vowel	ក	ា		កា
Consonant + consonant Shifter + vowel	ស	◌្	◌ី	ស្
Consonant + consonant Shifter + vowel	ប	◌	◌ី	ប្

Table 2: Shape and Position Deformation

III. RESEARCH GOALS AND RESEARCH FINDINGS

III.1. Development of Sustainable Human Resource Capacity for R&D

Before the existing of Khmer Unicode in Computer, Khmer Natural Language Processing in computer was poor. Khmer needs to learn a second foreign language such as French or English to adapt the IT world. Documents were preferably typed in English or French in most cases. It is difficult to edit, decorate, or check for grammar or spelling error in Khmer Documents which are typed in legacy fonts. Most of the documents are full of errors and bad formatted.

To resolve the facing problems, research study on Khmer language characteristics must be conducted. Khmer is deserved to do the research as s/he has the basic knowledge of his/her own speaking and writing language rather than foreigners. On the other hand, knowledge on NLP among Cambodian is still limited. The capacity building among the young Cambodian developers is done by let them learn to do rather than learn to use.

For instance, we invited a developer from Pakistan who masters in the NLP to teach and share his experiences among the young Cambodian developers to learn how to develop their assigned projects. In case one person is leaving, i.e. for their study, or another job, that person will have to teach the new comer all his experiences related to the projects. This leads to the development of sustainable Human Resource Capacity building for R&D among both female and male Cambodian fellows.

III.2. Culture and Policy Promotion for Local Language Computing

With the applications developed by Pan Localization Cambodia (PLC) such as, Khmer spell checker, Khmer OCR, Khmer Line Breaking, Khmer Word Segmentation, Khmer Sorting, Unicode fonts, and Khmer SMS, Khmer NLP in computer has been promoted. The difficulty in editing, or formatting the Khmer documents is overcome by using Unicode fonts, Khmer Spell checker, Khmer Line Breaking, Khmer sorting. This enhances the Khmer writing culture in ICT field. Rendering rule in Fonts enable users to type Khmer according to their spelling as shown in Table 2. Through some trainings and workshops in Institute of Technology of Cambodia, both students and teachers are welcome the Khmer Unicode in Computer because they understand the benefits of Khmer Unicode and the Khmer based application developed by PLC. We have also conducted the training in Ministry of Interior and National Assembly about the utilization of the package we developed. As a result, MoI and NA have accepted the Khmer Unicode and the Applications developed by PLC to be used internally. They are also continuously giving the feed back to us.

III.3. Measurement and Evaluation

The achievement of the Khmer application developed by PLC give a huge advantage among both end-users and developers. End-users are now happy to type the documents in Computer based on Khmer Unicode fonts because they can check for errors, sort, and format easier than using the legacy fonts. On the other hand, for developer, they improve their capacity on R&D especially in NLP. They also can share the younger promotion about their experience. With the experience gained from PLC they can both further their study and work in a higher level of R&D with a high responsibility.

III.4. Marketing and Dissemination

The awareness of the advantages of Khmer Unicode and Khmer based application built by PLC has been reached among the computers end-users and developers. The information seeker/provider can search/promote more Khmer contents in Computer or Internet. The developers are now able to localize the applications in Khmer to answer the requirements of local users.

IV. PROJECT OUTPUTS

IV.1. Research Reports

In this phase, we produced research reports as followings that can be found at www.panl10n.net:

- Khmer OCR Technical Report
- Khmer SMS Technical Report
- Khmer Encoding Report
- Khmer Collation Report
- Khmer Segmentation Report

IV.2. Research Publications

In this phase, we produced research publications as followings that can be found at www.panl10n.net:

- Khmer OCR for Limon R1, Size 22
- Noise Removal Using Connected Component Labeling
- Rendering, Keypad Layout and Input Method to Support Khmer Script on Mobile Devices

V. SUSTAINABILITY OF WORK AND FUTURE PROSPECT

V.1. Human Resource Capacity Building

To enable the sustainability of Human Resource Capacity Building, we build it among the Cambodian team so that they will be able to not only use the technology developed by others, but also resolve the problem they meet and share experience among the team. Sharing experience among team is the way that we can improve and ensure the sustainability of the Capacity Building i.e. one new resource must be overlapped with the leaving resource.

V.2. Adoption of Local Language Computing by End Users

The end-users now adapt well with Khmer Unicode contents. With to-be-released Khmer SMS, we hope that SMS in Khmer will be very popular among Cambodian People who can at least read and write Khmer.

V.3. Influence on Language Computing Policy

More applications and contents are now available in Khmer. Khmer now prefer produce the content and develop the software in Khmer. We expect that more and more Khmer-based applications will be more and more produce with the help of the Unicode and the applications developed by PLC

V.4. Beyond PAN Phase II

Khmer based applications have made a big change in ICT field in Cambodia. End-users enjoy the application and produce contents in Unicode. However, the Unicode fonts have not been supported all the application yet. Some rules are missing in some application like Adobe Photoshop and Microsoft Visio. Khmer Spell Checker has also limit capacity in checking errors because the small corpus. Khmer OCR is able to recognize only two fonts. We are trying our best to reduce the bugs and improve the capacity of the developed applications. We are also seeking for grant to ensure the sustainability of the on-going projects.